NIST / RT-2002 workshop
PSTL

Patrick Nguyen

Luca Rigazio

Yvonne Moh

Jean-Claude Junqua

Plan

- Who we are
- LVCSR in PSTL
- Meta-data systems
- STT systems
- Distinctive features
- Conclusion

Panasonic Speech Technology Laboratory (PSTL)

of

Panasonic Technology Company (PT), a division of Matsushita Electric Company of America (MECA)

- PT has more than 10 research labs in the USA
- About 20 researchers in PSTL
- Synthesis and recognition
- Contributions:
  - PLP, eigenvoices, modified Jacobian adaptation, Lombard effect, speaker adaptation
  - Focus: small vocabulary , noise-robustness, medium-vocabulary => small foot print (hardware)

LVCSR in PSTL

- About 2 years-old
- Small team, scarce resources
- Includes decoder, training, and language modeling
- Written entirely from scratch

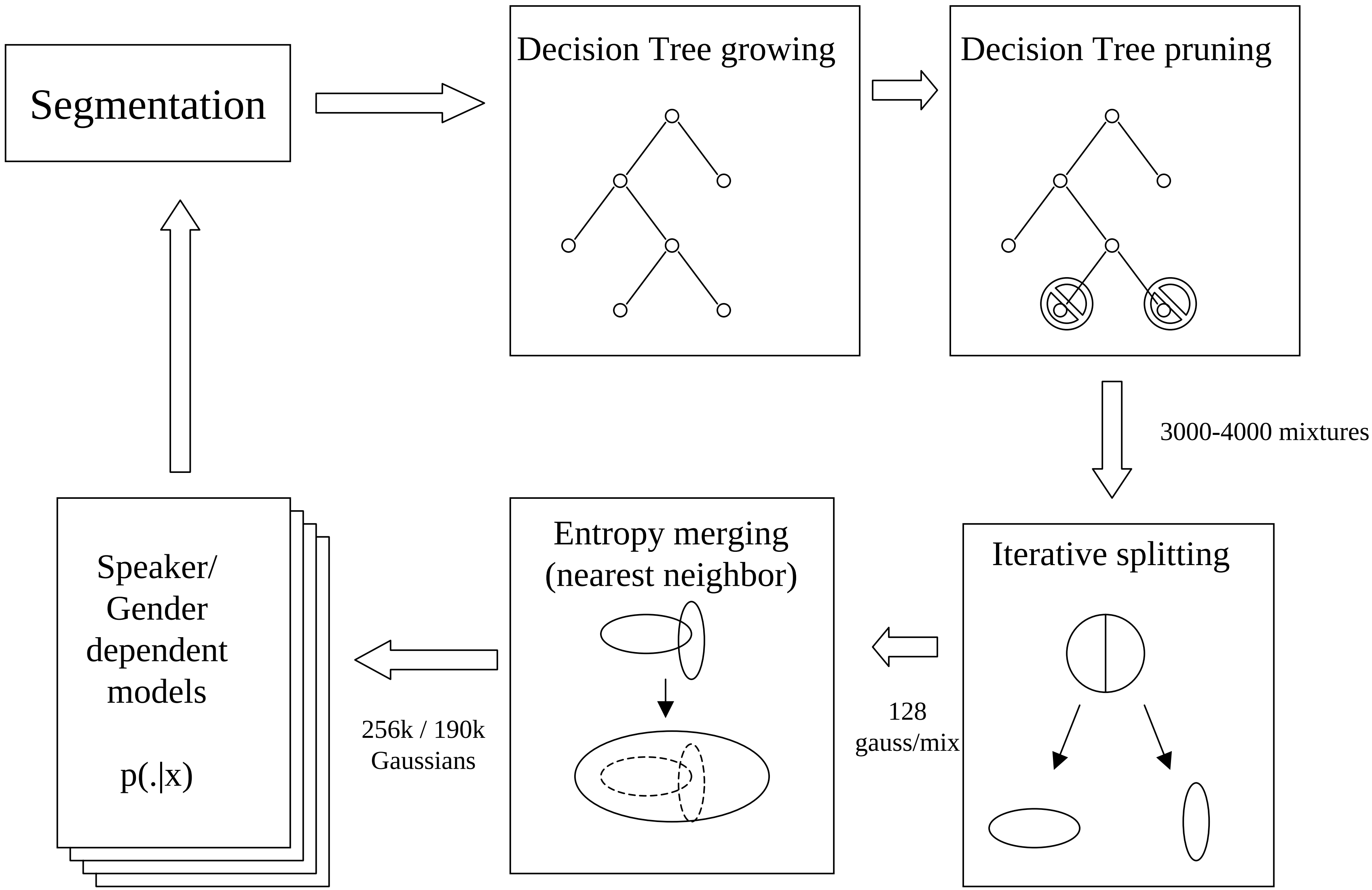
Disclaimer

- Most features are standard => will not describe them
- Emphasis on “distinctive” features
- Please refer to system descriptions or paper for details

Meta-data systems

- Same except for the parameterization
- Generate condition-dependent segmentation
- Oversegment and merge contiguous cuts
- BIC criterion to segment
- BIC criterion to cluster
- Designed for regression (speaker adaptation) and not classification

Speech-To-Text (STT) systems: Training

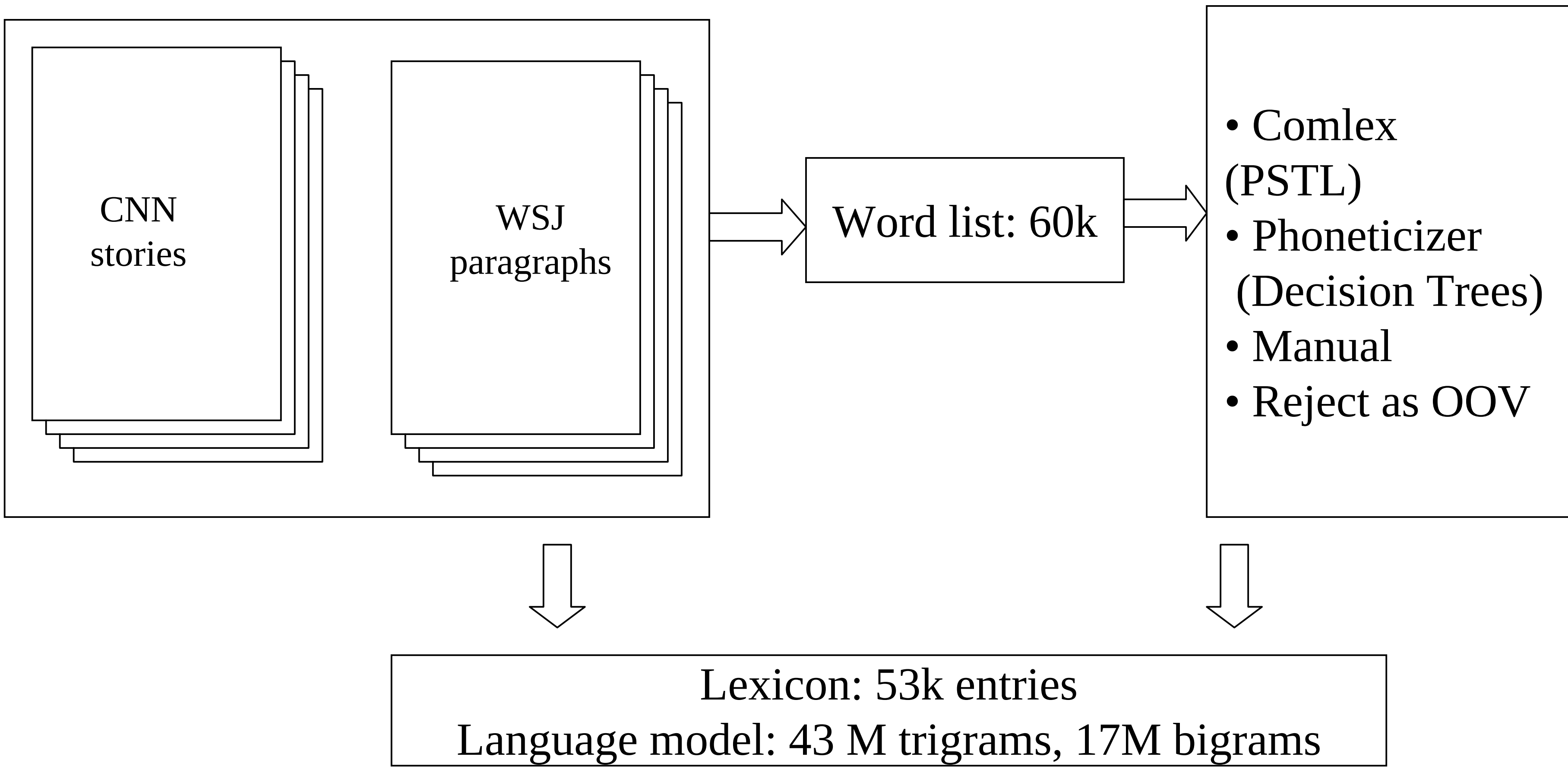


EWAVES: decoder

- Single state-based lexical tree
- Viterbi trigram topology
- Bigram lookahead (cached)
- Histogram pruning
- Very fast word-internal decoding
  (10M state hypotheses / sec)
- Parallel architecture

STT bnews system: language modeling

- (SWB: provided by A. Stolcke from SRI)



Some distinctive features

- EWAVES (optimized Viterbi)
- word-internal (xwrd through compounding)
- training procedure

- low-level optimization: hcache
- iconic adaptation

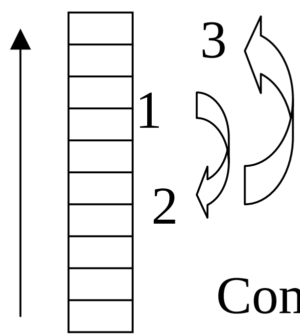
Horizontal Caching: compute more, go faster

- SWB: 70% distributions are used, 35xRT => 14xRT
- WSJ: 40% distributions are used, 4xRT => 2xRT
- Optimize memory fetching: fetch 1 distrib / one block of observations

On-demand computing

For t=1…T,
 for all active states,
  compute mixture M of state

Mixture #



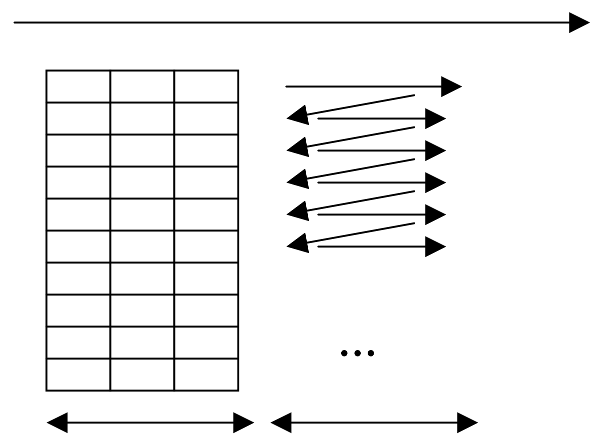
Compute order

Likelihood table for time a time t

Horizontal Caching (frame blocking)

For b=1…T/10,
 Compute all distributions for block b

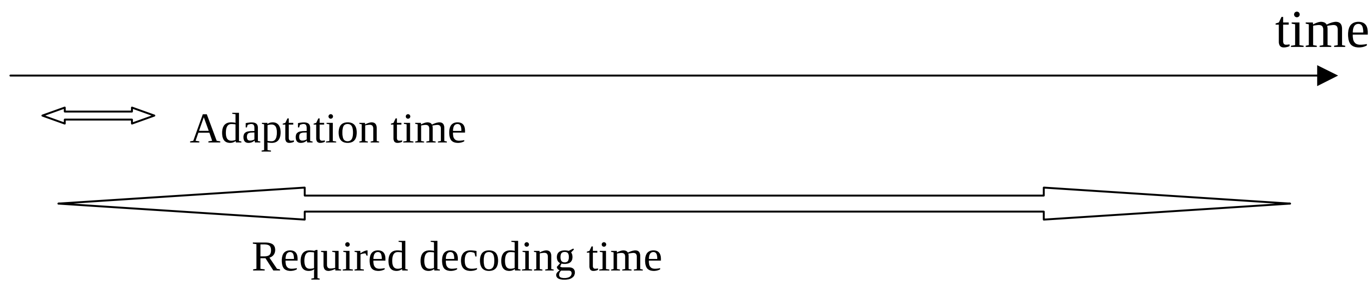
Time t



Block 1 Block 2

Iconic adaptation / almost 1-pass system

- Adaptation takes time => use very-fast adaptation
- Adaptation time:
  - First-pass decode O(time)
  - Accumulation O(time)
  - Compute transformation O(model)
  - (a) Model update O(model)
  - (b) Feature update O(time)
- => Reduce time of first-pass decode, accum, and feature update within adaptation



- Why adaptation? Faster 2nd-pass + lower WER

Two Features under construction

- Construction of model-space constraints

- Linear transformation of feature spaces

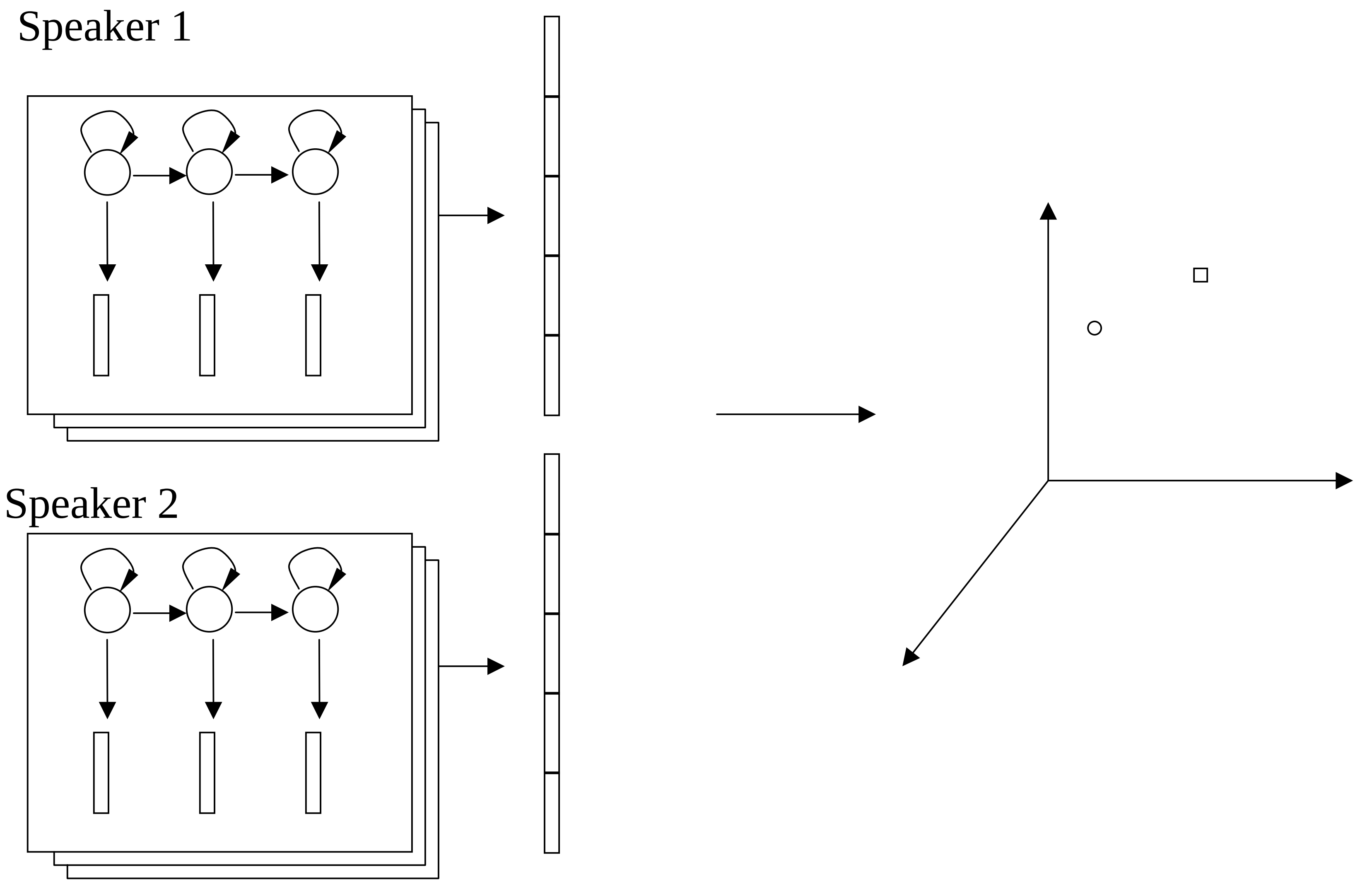
Construction of Model-Space Constraints

- Goal:
  - Model HMM parameters as random variables
  - Encode observed speakers' patterns
  - Generalization of gender-dependent modeling
- Solution:
  - Assume normal & isotropic HMM models
    => eigenvoices (well-known)
  - Assume piecewise normality
  - Dimension reduction according to HMM source
  - Use for discriminative purposes

Eigenvoices

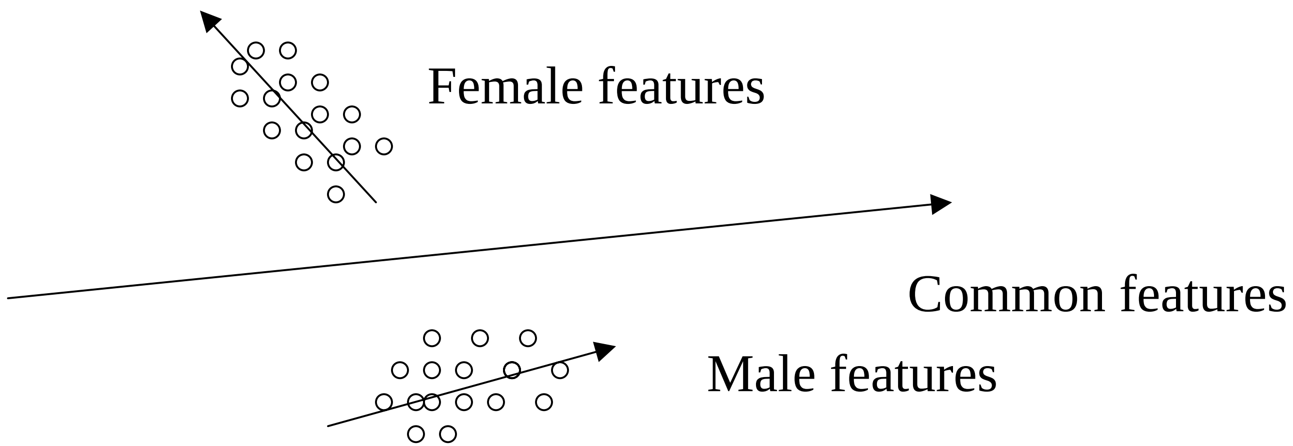
- Build speaker-dependent models
- Observe their distribution
- Assume joint Gaussianity of parameters
- Given normal assumption, find Principal Components of autocovariance matrix
- => Linear space spans possible models
- => Use for rapid speaker-adaptation

Eigenvoices: model parameters are random variables



Piecewise normality
(ICASSP2002)

- Piecewise normal spaces
- Simplest generalization of non-linearity
- Dependency is a function of position
- E. g.: gender-dependent eigenspaces



Discriminative eigenspaces
(ASRU2002)

- Eigenvoices: kind of multi-dimensional SAT (=> CAT)
- MMIE: discriminative, but SI discrimination is suspect

- Combine both => Discriminative adaptation using prior speaker information
- Criterion matrix: J = H - X,
  - H = ML components (standard)
  - X = MMIE components (gradient due to errors)

Feature-space adaptation
?ICSLP-2002?

- Transform o => Ao + b
- No closed-form solution for full-matrix A
  – Use fast numerical method (Gales/tr291)
  – Use A = U D, then find unitary U, then D (EM)
  – Our solution: A = L U
    - L, U are lower and upper-triangular matrices
    - Closed-form solution for L, then U (EM)
    - MAP solution is available (modified Rayleigh-Maxwell distribution)

Dev vs Eval results

- SWBD1: (le10xRT)
  - Manual: eval00: 33% WER / eval02: 37% WER
  - Auto: eval00: 33% WER / eval02: 36% WER
  - Typical (estimate): 30% WER / eval02: 22% WER
- SWB harder than our dev set (could be overtuning)
- (auto has a larger beam)

- BN:
  - 10xRT: eval98: 22%WER / eval02: 20% WER
  - 1xRT: eval98: 27% WER / eval02: 24% WER
- BN easier than our dev set

SWB results

- Pre-evaluation estimation
  - PSTL: 34% WER @ 10xRT
  - 1st pass: 28-35% WER @ 20-40xRT
  - Worst: 37% WER @ unlimited RT
- Post-evaluation (SWBD1)
  - PSTL: 37% WER
  - AT&T le1xRT: 29.5% WER
  - CUHTk-late: 22.3% WER

BN results

- Started working on BN in January 2002
- Built Speaker Segmentation / Clustering
- Trained acoustic models
- Trained LM

- Comparatively good results

Conclusion

- PSTL characteristics
  - exploratory conditions
  - most MD and STT systems (4 STT + 2 MD systems)
  - portable system (BN built last minute)
  - more experience in speaker adaptation

- => More work on baseline
- => More distinctive features